DOCUMENT RESUME

ED 371 006 TM 021 637

AUTHOR Sadek, Ramses F.; Huberty, Carl J.
TITLE Using Monte Carlo Studies in Discriminant Analysis: An Overview.
PUB DATE Apr 94
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Discriminant Analysis; *Estimation (Mathematics); *Evaluation Methods; Foreign Countries; Literature Reviews; *Monte Carlo Methods; *Research Methodology; Simulation; Statistical Studies
IDENTIFIERS Outliers; Variables

ABSTRACT

This study presents an overview of Monte Carlo studies in discriminant analysis. Some common questions about the use of Monte Carlo techniques are answered through a brief literature review of articles on discriminant analysis in which Monte Carlo methods are used. The articles cover many research points, such as comparing error rate estimates, evaluating different discriminant rules, and studying outlier influence. The study may be of assistance to researchers who are interested in conducting Monte Carlo studies, especially in choosing the values of the parameters of factors under consideration. Recommendations are made for the choices of discriminant analysis parameters in a Monte Carlo study. Five tables present results from the literature review. (Contains 36 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

USING MONTE CARLO STUDIES IN DISCRIMINANT ANALYSIS

"AN OVERVIEW"

RAMSES F. SADEK[1] & CARL J HUBERTY[2]

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RAMSES F. SADEK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 371 006

Abstract

To conduct a Monte Carlo study in discriminant analysis, one is faced with many questions. What is the value of p, the number of variables in each population? What are, μ_k, the mean vectors? What are the covariance matrices, Σ_k? Are they equal? What is the degree of separation between populations, Δ_{ij}? What are the sample sizes, n_k? How many replications are needed?

The purpose of the current study is to present an overview of the Monte carlo studies in discriminant analysis. We will try to give examples of answers to the above questions through a brief literature review of the topic. The study contains a review of articles on the topic of discriminant analysis in which Monte Carlo sampling techniques were used. The articles covered many research points such as comparing error rate estimates, evaluating different discriminant rules, and studying the outlier influence. The study may be of assistance to researchers who are interested in conducting Monte Carlo studies, especially in the process of choosing the values of the parameters or factors under consideration.

KEY WORDS: Discriminant Analysis, Monte Carlo, Simulation.

1. INTRODUCTION

Naylor, Balintfy, Burdick, and Chu (1966, p. 3), defined simulation as a numerical technique used to conduct experiments on a digital computer, which involves certain types of mathematical and logical models that describe the behavior of a system over extended periods of time. It is a technique of performing sampling experiments on a model of a system. Naylor et al. (pp. 5-7) described many situations in which simulation can be successfully used.

---

[1]Assistant Professor, department of Statistics, Faculty of Economics & Political Sciences, Cairo University, EGYPT.

[2]Head, department of Educational Psychology, University of Georgia, USA.

TM021637

BEST COPY AVAILABLE

Examples of such situations are: (a) It may be either impossible or very expensive to obtain data from certain phenomena in the real world; (b) The observed system may be too complex to be described in terms of a set of mathematical equations for which the analytic solutions are obtainable; (c) Even if the mathematical model can be formulated to describe some system of interest, it may not be possible to obtain a solution to the model by straightforward analytic techniques; and (d) It may be impossible or too expensive to validate experiments on the mathematical model describing the system.

Naylor et al. (1966, pp. 8-9) listed some reasons for using simulation analysis:

1) Simulation can be used to experiment with new situations about which we have little or no information so as to prepare for what may happen.

2) Detailed observation of the system being simulated may lead to a better understanding of the system and to suggestions for improving it, suggestions that otherwise would not be apparent.

3) The experience of designing a computer simulation model may be more valuable than the actual simulation itself. For example, the knowledge obtained in designing a simulation study may suggest changes in the system being simulated. The effects of these changes can then be tested via simulation before implementing them on the actual system.

4) Through simulation, the effects of certain informational, organizational, and environmental changes on the operation of a system could be studied, by imposing some alterations in the model

of the system and observing the effects of these alterations on the system's behavior.

5) Simulation makes it possible to study and experiment with the complex internal interactions of a given system whether it be a firm, an industry, an economy, or some subsystem of one of the these.

6) Simulation can be used as a pedagogical device for teaching both students and practitioners basic skills in theoretical analysis, statistical analysis, and decision making.

7) Operational gaming has been found to be an excellent means of stimulating interest and understanding on the part of the participant, and in particular useful in the orientation of persons who are experienced in the subject of the game.

8) Simulation of complex systems can yield valuable insights into which variables are more important than others in the system and how these variables interact.

9) Simulation can serve as a "pre-service test" to try out new policies and decision rules for operating a system, before running the risk of experimenting on the real system.

10) Simulations are sometimes valuable in that they afford a convenient way of breaking down a complicated system into subsystems, such of which may then be modeled by an analyst or team that is expert in that area.

11) Simulation makes it possible to study dynamic systems in either real time, compressed time, or expanded time.

12) When new components are introduced into the system, simulation can be used to help foresee bottlenecks and other problems that may arise in the operation of the system.

Computer simulation also enables us to replicate an experiment. Replication means rerunning an experiment with selected changes in parameters or operating conditions being made by the investigator. Simulation is indeed an invaluable and very versatile tool in those problems where analytic techniques are inadequate. However, it provides only statistical estimates rather than exact results, and compares alternatives rather than generating the optimal one.

Because sampling from a particular distribution involves the use of random numbers, a simulation study is sometimes called a Monte Carlo study. The term "Monte Carlo" was introduced during World War II, as a code word for the secret work at Los Alamos; it was suggested by the gambling casinos at the city of Monte Carlo in Monaco. The Monte Carlo method was then applied to problems related to the atomic bomb.

Monte Carlo methods have been used for evaluating multidimensional integral and differential equations, and for simulating some parameters of queues and networks. Sampling random variates from probability distributions is another field of Monte Carlo applications. Perhaps, the Monte Carlo method is now the most powerful and commonly used technique for analyzing complex problems.

Rubinstien (1981, p. 12) listed three differences between the Monte Carlo method and simulation:

1. In the Monte Carlo method time does not play as substantial a role as it does in stochastic simulation.

2. The observations in the Monte Carlo method, as a rule, are independent. In simulation, however, we experiment with the model over time so, as a rule, the observations are serially correlated.

3. In the Monte Carlo method it is possible to express the response as a rather simple function of the stochastic input variate. In simulation, the response is usually a very complicated one and can be expressed explicitly only by the computer program itself.

2. MONTE CARLO STUDIES AND DISCRIMINANT ANALYSIS

Three of the main reasons for conducting a Monte Carlo study in discriminant analysis are: a) The observed system may be too complex to be described in terms of a set of mathematical equations for which the analytical solutions are obtainable; b) It may not be possible to obtain a solution to the model by straightforward analytical techniques; and c) It may be impossible or too expensive to validate experiments on mathematical model describing the system. These three reasons are valid for the areas of: 1) Studying outlier detection and influence; 2) Comparing error rate estimates; 3) Evaluating discriminant rules; and 4) Studying a discriminant rule behavior under nonoptimal conditions.

In general, the main steps to conduct a Monte Carlo study in discriminant analysis are as follows:

1. Specify the parameters of interest such as the number of variables (p), the mean vectors (μ_k, k = 1,...,K), the covariance matrices (Σ_k, k = 1,...,K) and/or the degree of overlap/separation between the groups (in terms of the Mahalanobis distance or the expected error rate).

2. From each population G_k, sample n vectors of size p x 1, with the mean vector μ_k, and covariance matrix Σ_k, specified in the previous step.

3. Use the vectors from step 2 to construct a discriminant rule.

4. Using the discriminant rule from step 3, calculate some statistics of interest (e.g., the hit rates).

5. Replicate steps 2 through 4 many times, and calculate the averages and the standard deviations and any other statistics or estimates of interest.

6. Analyze and interpret the results by graphical or inferential methods.

There are three distinct methods of performing the six steps specified above. The main difference among these three methods is concerned with the second step mentioned above, namely, generating a multivariate vector from a population with some parameters.

The first and most popular method is to generate samples from populations with specified parameters using a random number generator. This method is usually referred to as Monte Carlo sampling, Monte Carlo simulation, or simulation experiment (see for example, Hora & Wilcox, 1982, Remme, Habbema & Hermans, 1980, Sadek & Huberty, 1992, and Snapinn & Knoke, 1989).

In the second method, populations G_k ($k = 1,\ldots,K$) of sizes N_k with the specified parameters are generated. Then, samples of sizes n_k are selected with replacement from that population (each multivariate observation has a probability $1/N_k$ of being represented in the sample). The main problem here is that some observations may be represented more than once in the same sample. This will depend on the population size, N_k. Another concern with this method is the choice of the population sizes (N_k) which are supposed to be infinite. The question here is how large N_k should be

considered infinite? This method is known as the resampling technique or bootstrapping (as examples, see Chatterjee & Chatterjee, 1983, Efron, 1983, and Freed & Glover, 1986).

For the third method, real data sets are used to represent populations. Then, with replacement, samples are drawn from this data set. In this case, the population parameters are calculated from the data set. A good example of this procedure is given in Huberty, Wisenbaker, and Smith (1987). Two main problems with this approach are that a real data set has a relatively small size, and represents a unique situation.

Regardless of the method used, to conduct a Monte Carlo study in discriminant analysis, one is faced with many questions. What is the value of p, the number of variables in each population? What are, μ_k, the mean vectors? What are the covariance matrices, Σ_k? Are they equal? What is the degree of separation between populations, Δ_{ij}? What are the sample sizes, n_k? How many replication are needed?

3. CHOICE OF FACTORS FOR A MONTE CARLO STUDY IN DISCRIMINANT ANALYSIS

The purpose of the current study is to present an overview of the Monte carlo studies in discriminant analysis. This overview would help one to make reasonable choices of the value(s) of the parameters to be used in a Monte Carlo study in a discriminant analysis. A review of some published articles on the topic of discriminant analysis in which Monte Carlo sampling techniques were used was conducted. The articles covered many research points such as comparing error rate estimates, evaluating different discriminant

rules, investigating variable selection methods, and studying the outlier influence. Most of these articles were published in various journals such as Communications in Statistics, Computational Statistics & Data Analysis, Decision Sciences, Journal of American Statistical Association, Multivariate Behavioral Research and other statistical and statistics-related journals.

In the current section, a discussion of the choice of the discriminant analysis parameters values is given. Subsection 3.1, deals with the choices of the populations parameters such as the mean vectors, μ_k, the covariance matrices, Σ_k, and Mahalanobis distance between the groups, Δ_{ij}. Subsection 3.2 deals with the choices of number of observations in each group (sample sizes), the number of variables in each population (dimensionality) and the number of groups, K. Subsection 3.3 deals with the number of replication in a Monte Carlo study.

3.1 Choice of μ_k, Σ_k, and Δ_{ij}

The mean vectors, μ_k, covariance matrices, Σ_k, and Mahalanobis distance, Δ_{ij}, are related through the relationship

$\Delta_{ij}^2 = (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)$. Then, determining two of the above three terms implies the determination of the third term. It is common to study the effect of Δ_{ij} when one studies the outlier influence, compares error rate estimates or evaluates a discriminant rule. The effect of hetrogeniety of the covariance matrices is usually investigated when one evaluates a discriminant rule based on data from normal distribution.

The most important factor here is Δ_{ij}, the degree of overlap/ separation between groups i,j (i,j= 1,..,K) in terms of Mahalanobis distance. Two alternatives are available.

First, choose the covariance matrix and the mean vectors, then the value of Mahalanobis distance is calculated from the above equation. Examples of this are found in Chatterjee & Chatterjee (1983), Friedman (1989), Joachimsthaler & Stam (1988), Koehler & Erenguc (1990), Lahiff & Whitcomb (1990), McLachlan (1980), Moore, Whitsitt & Landgrebe (1976), Raveh (1989), Remme, Habbema & Hermans (1980), Stam & Ragsdale (1992) and Young et al. (1986). This approach is typically used if the degree of overlap is not a factor to be studied. In mathematical programming approach the degree of overlap may be specified in a different way rather than Mahalanobis distance. Examples of this kind are found in Freed & Glover (1986) and Rubin (1990).

Second, choose the values of Mahalanobis distance and the covariance matrix, then the mean vectors could be determined from the above relationship. Here, the degree of overlap/separation between the groups is a factor to be studied. The choice of the mean vectors and the covariance matrix are selected to produce some specific values of Δ.

Usually various values of Δ are chosen such that they represent small, moderate, and large overlaps. However, the number of chosen values for Δ varies from one study to the other. For example, Joachimsthaler & Stam (1988) and Koehler & Erenguc (1990) used one value of 1 for Δ. The values of Δ = 1, 2, and 3 were chosen by Lachenbruch, Sneeringer, & Revo (1973), the values of .5, 1, 1.5 were chosen by Sadek & Huberty (1992), and 14 values of Δ between

0 and 4 were chosen by Glick (1978). Some examples are summarized in Table 1 below.

Table 1: Examples of choices of Mahalanobis distance, Δ

Authors	Chosen values of Δ
Dorans (1988)	1, $\sqrt{2}$, 2, $\sqrt{8}$
Ganesalingam & McLachlan (1980)	1, 1.5, 2, 3
Ganeshanandam & Krzanowski (1989)	1.01, 2.53
Glick (1978)	0, .1, .2, .4, .5, .6, .8, 1, 1.25, 2(.5)4
Greene & Rayens (1989)	1.5, 2, 3, 5, 5
Hora & Wilcox (1982)	1, 2
Lachenbruch, Sneeringer & Revo (73)	1, 2, 3
Lesaffre, Willems & Albert (1989)	$\sqrt{.5}$, 1, $\sqrt{2}$, 2
McLachlan (1980)	1, 2
Sadek & Huberty (1992)	.5, 1, 1.5
Snapinn & Knoke (1984 & 1985)	0,.5, 1, 1.5, 2, 2.5, 3
Snapinn & Knoke (1989)	0, 1, 2, 3
Young, Marco & Odell (1986)	1.01, 1.31, 1.55

If the two populations are multivariate normal with equal priors, there is a relationship between the expected misclassification rates and Mahalanobis distance. Mahalanobis distance is selected so that a desired misclassification rate is attained (for example, a Mahalanobis distance value of 3.29 would produce a misclassification rate of .10). Examples of the application of this idea are found in Bayne et al. (1983), Lachenbruch & Mickey (1968), Page (1985), and Vlachonikolis (1986). A summary of these choices is given in table 2 below.

Table 2: Examples of choices of Error rates

Authors	Chosen values of error rate
Bayne et al. (1983)	.05, .15, .25, .35,.45
Lachenbruch & Mickey (1968)	.05, .1, .15, .20, .25, .30
Page (1985)	.10, .20, .30
Remme, Habbema, & Hermans (1980)	.10
Sadek & Huberty (1992)	.15, .30, .45
Vlachonikolis (1986)	.30, .50, .70

The identity matrix, $\mathbf{I}$, is the most common choice for the covariance matrix. For example, see Greene & Rayens (1989), Joachimsthaler & Stam (1988), Koehler & Erenguc (1990), Lachenbruch & Mickey (1968), Lahiff & Whitcomb (1990), McLachlan (1980), Page (1985) and Sadek & Huberty (1992).

Different forms of Σ_k (instead of a common Σ) are required if the effect of hetrogeniety of the covariance matrices is under investigation. Usually this takes place when one compares discriminant rules that includes the quadratic discriminant rule. Using different forms of Σ_k are also found when the researchers use real life data. Examples of the first situation are found in Freed & Glover, 1986 ($\Sigma_2 = \Sigma_1, 9\Sigma_1$), Friedman, 1989 ($\Sigma_k = k.\Sigma_1$, $k = 1,2,3$), Greene & Rayens, 1989 ($\Sigma_1 = \mathbf{I}$, $\Sigma_2 = (1/32).\mathbf{I}, (1/8).\mathbf{I}, (1/2).\mathbf{I}$), Joachimsthaler & Stam, 1988 ($\Sigma_2 = \Sigma_1, 2\Sigma_1, 4\Sigma_1$), Koehler & Erenguc, 1990 ($\Sigma_1 = \mathbf{I}$, $\Sigma_2 = \Sigma_1, 2\Sigma_1, 4\Sigma_1$), Raveh, 1989 ($\Sigma_2 = \Sigma_1, 3\Sigma_1, 4\Sigma_1$), Remme, Habbema, and Hermans, 1980 ($\Sigma_2 = \Sigma_1, 2\Sigma_1, 4\Sigma_1, 16\Sigma_1$), and Rubin, 1990 ($\Sigma_2 = \Sigma_1, 9\Sigma_1$).

After selection of the covariance matrix, the mean vectors which produce some values of Mahalanobis distance must be determined. Two ways of determining the shape of the mean vector are found in the literature. The first one is to choose two vectors in such a way that the difference between the two vectors is spread among all variables, so that each element in the first vector differs from the corresponding element in the second vector by a constant (e.g., $\mu^T_1 = (2,2,2,2)$, $\mu^T_2 = (1.5,1.5,1.5,1.5)$. Examples of that choice are found in Ganesalingam & Mclachlan (1989), Joachimsthaler & Stam (1988), Koehler & Erenguc (1990), Lahiff & Whitcomb (1990), Raveh (1989), and Snapinn & Knoke (1989). If the covariance matrix is

the identity matrix, as a general rule, μ_i and μ_j may be chosen so that $\mu^T_j = \mu^T_i \pm (\Delta/\sqrt{p})(1,\ldots,1)$.

The second one is to choose two vectors differing by only one element (e.g., $\mu^T_i = (0,0,0,0)$, $\mu^T_j = (2,0,0,0)$). Examples of that can be found in Friedman (1989, Lachenbruch, Sneeringer, & Revo (1973), Lachenbruch & Mickey (1968), Page (1985), Remme, Habbema, and Hermans (1980) and Snapinn & Knoke (1984). If the covariance matrix is the identity matrix, then, the choice of $\mu^T_j = \mu^T_i \pm (\Delta,0,\ldots,0)$ produces the required squared Mahalanobis distance, Δ^2.

3.2 Choice of p, K, and n_k

When the study involves a number of dimensions, various values of p are to be considered, otherwise, only one value of p is considered. It is up to the researcher to determine the value (s) of p which he/she will take into consideration. One value of p was considered by various authors. Chatterjee & Chatterjee (1983), Glick (1978), and Moore, Whitsitt, & Landgrebe used only p = 1. Bayne at al. (1983), Freed & Glover (1986) Fukunaga & Kessell (1971), and Rubin (1990) used p = 2. Joachimsthaler & Stam (1988), Koehler & Erenguc (1990), and Stam & Ragsdale (1992) used p = 3. Young et al. (1986) used p = 6, Dorans (1988) used p = 8, and Ganeshanandam & Krzanowski (1989) used p = 10.

Two values of p were considered by Hora & Wilcox, 1982 (p = 5, 10), Greene & Rayens, 1989 (p= 4, 10), Lesaffre, Willems, & Albert, 1989 (p = 3, 5), Raveh, 1989 (p = 2, 3), and Snapinn & Knoke, 1989 (p= 10, 20). Three values of p were chosen by Ganesalingam & McLachlan, 1980 (p= 1, 2, 4), Page, 1985 (p= 4, 8, 20), Remme, Habbema, & Hermans, 1980 (p = 2, 6, 10), Sadek & Huberty, 1992 (p= 2, 4, 6),

Snapinn & Knoke, 1984, Snapinn & Knoke 1985, and Vlachonikolis, 1986 (p = 1, 3, 5). More than three values of p were considered by Friedman, 1989 (p = 6, 10, 20, 40), Lachenbruch & Mickey, 1968 (p = 2, 4, 8, 20).

Equal group sizes was a common choice for most of the studies. Lachenbruch & Mickey (1968) suggested that $n_k \geq 3$ is necessary to obtain a reasonable hit rate estimate. Huberty, Wisenbaker, & Smith (1987) recommended that a minimum n_k/p ratio of approximately 3 might be considered a definition of a large group size. Most of the reviewed papers satisfied the condition of Huberty, Wisenbaker & Smith. Exceptions are found in some data conditions of Friedman (1989), Greene & Rayens (1989), Lachenbruch & Mickey (1968), Page (1985), Remme, Habbema, & Hermans (1980).

Around one third of the reviewed papers considered one value for group size. Glick (1978) considered a group size of 10, Lahif & Whitcomb used a group size of 16, Snapinn & Knoke (1984, 1985) considered a group size of 25, while 50 observations per group were used by Freed & Glover (1986), Koehler & Erenguc (1990), Raveh (1989), and Rubin (1990).

When the effect of the sample size was under consideration, researchers used different sample sizes. Two different ways of selecting the group sizes were found in the literature. The first was choice of n_k as a multiple of p. The second, was direct determination of different values of n_k. Examples of the former are found in Hora & Wilcox, 1982 (n_k = 3p, 10p), Lesaffre & Willems, 1989 (n_k = 3p, 5p, 10p), and Huberty & Sadek, 1992 (n_k= 3p, 6p, 9p). Examples of the latter are found many articles. Two choices of n_k were considered by Friedman, 1989 (n_k = 10, 20), Ganeshanandam &

4. SUMMARY AND CONCLUSIONS

A brief summery of various authors' choices of K (number of groups), p (number of variables), n_k (group size), group separation, and number of replications is given in Table 5. Our recommendations for the choices of the discriminant analysis parameters in a Monte Carlo study are as follows:

1. Three values of Δ would be sufficient to represent low, moderate, and high separation between groups. If applicable, the expected error rate should be used to determine the desired group separation levels.

2. If the effect of dimensionality is desired, different values of p between 1 and 10 is suggested. In variable selection studied, a larger number of variables may be desired.

3. Using n_k as a multiple of p is preferable. Values of n_k between 3p and 10p are recommended.

4. Concerning number of replications, Sadek & Huberty (1992) reported that gain one may obtain by increasing the number of iteration from 2000 to 5000 is negligible. Because of the availlability of computing facilities, one may get the advantage of running a large number of iterations. The authors think that a number of replication between 500 and 2000 would be sufficient.

5. Concerning the covariance matrices, no loss of generality occurs when $\Sigma = I$, the identity matrix. Since the linear discriminant function can be thought of as first reducing the covariance matrix to its principal components and then standardizing so that the variation is equal in all directions. A similar justification is given by Hand (1981, p. 139). For heterogeneous covariance matrices $\Sigma_k = m.I$, where m is an integer, is recommended.

Krzanowski, 1989 (n_k = 25, 45), Ganeshanandam & McLachlan, 1980 (n_k= 20, 50), McLachlan, 1980 (n_k=20, 20), and Remme, Habbema & Hermans, 1980 (n_k = 15, 35). Three choices of n_k were considered by Bayne, 1983 (n_k = 25, 50, 100), Chatterjee & Chatterjee, 1983 (n_k = 10, 20, 50), Fukunaga & Kessell, 1971 (n_k= 100, 200, 400), Snapinn & Knoke, 1984 (n_k = 10, 20, 25), Stam & Ragsdale, 1992 (n_k = 25, 50, 100), and Vlachonikolis, 1986 (n_k = 50, 100, 200). The largest number of choices for n_k was used by Moore, Whitsitt & Landgrebe (1976) where they used ten different values of n_k between 20 & 200 with an increment of 20.

3.3 The Choice Of Number Of Replications

Most Monte Carlo experiments involve replicating the experiment for a number of times. The number of replications varies from one study to another. In the area of discriminant analysis, the number of replications varied from 2 to 5000. A summary of some choices is given in table 3 below. The number of replications may vary in the same study. For example, Bayne et al., (1983) used 50000/n_k replications, n_k = 25, 50, and 100. Glick (1978) used 2000 and 4000 replications.

Table 3: Example of choices of number of replications

Replications	Authors
2	Lachenbruch & Mickey, 1968.
25	Remme, Habbema, & Hermans, 1980.
40	Fukunaga & Kessell, 1971.
50	Page, 1985, Stam & Ragsdale, 1992, Vlachonikolis, 1986.
100	Friedman, 1989, Joachimsthaler & Stam, 1988. Mclachlan, 1980, Raveh, 1989, Rubin, 1990, Snapinn and Knoke, 1989, and Young et al., 1986.
150	Dorans, 1988, and Greene & Rayens, 1989.
500	Hora & Wilcox, 1982, Lesaffre & Willems, 1989, Moore, Whitsitt, & Landgrebe, 1976.
1000	Freed & Glover, 1986.
2000	Glick, 1978, and Sadek & Huberty, 1992.
5000	Snapinn & Knoke (1984 & 1985).

Table 5: Summary

Author(s)	Notes	K[1]	p[2]	n_k[3]	Dist.[4]	Rep.[5]
1. Bayne et al., 1983	Comparing rules	2	2	25, 50, 100	.05(.10) .45*	50000/ n_k
2. Chatterjee & Chatterjee, 83	Error rate, bootstrap	2	1	10,20,50	2, 3	not given
3. Dorans, 1988	Error rate estimates	2	8	20, 40, 80, 160	1, √2, 2, √8	150
4. Freed & Glover, 1986	Evaluating rules (MP)	2	2	50	.5, .75**	1000
5. Friedman, 89	Comparing rules	3	6, 16 20,40	$n_1+n_2+n_3$ = 40	**	100
6. Fukunaga & Kessell, 1971	Error rate estimation	2	8	100,200, 400	?	40
7. Ganesalingam & McLachlan, 1980	Error rate estimates	2	1,2,4	20, 50	1, 1.5, 2, 3	100
8. Ganeshanandam & Krzanowski, 89	Variable selection	2	10	25, 45	1.01, 2.53	50
9. Glick, 1978	Error rate estimates	2	1	10	14 values from 0: 4	2000, 4000
10. Greene & Rayens, 1989	Comparing rules	4,8	4,10	6,9,31 13,21,51	1.5, 2, 3, 4, 5	150
11. Hora & Wilcox, 1982	Error rate estimates	3	5, 10	3p, 10p	1, 2	500
12. Joachimsthaler & Stam, 1988	Comparing rules (MP)	2	3	50	1	100
13. Koehler & Erenguc, 1990	Error rates	2	3	50	1	100
14. Lachenbruch & Mickey, 1968	Error rate estimates	2	2, 4, 8, 20	n_1=4:25, $n_2=n_1$, $2n_1, 3n_1$	.05(.05) .30*	2
15. Lahiff & Whitcomb, 1990	Error rates & outliers	2	2	16	1.06,2.12 2.83,4.24	N/A
16.Lesaffre, Willems & Albert,89	Error rate estimates	2	3, 5	3p,5p, 10p	√.5, 1, √2, 2	500

Author(s)	Notes	K[1]	p[2]	n_k[3]	Dist.[4]	Rep.[5]
17. McLachlan, 1980	Error rate, bootstrap	2	2, 4	5p, 10p	1, 2	100
18.Moore, Whitsitt & Landgrebe, 76	Hit rates	3	1	20(20) 200	.5, 2.75, 3.5	500
19. Page, 1985	Error rates	2	4, 8, 20	10,20 20,40 25,50	.10, .20, .30*	50
20. Raveh, 1989	Discriminant rules	2	2, 3	50	**	100
21. Remme, Habbema & Hermans, 1980	Comparing rules	2	2: 10	15, 35	.10*	25
22. Rubin, 1990	Comparing rules:MP,LD	2	2	50	.5, .75**	100
23. Sadek & Huberty, 1992	Outliers	2	2,4,6	3p,6p, 9p	.5,1, 1.5	2000
24. Snapinn & Knoke, 1984	Error rates estimators	2	1,3,5	10,20, 25	0, .5, 1, 1.5, 2, 3	5000
25. Snapinn & Knoke, 1985	Error rates	2	1,3,5	25	0(.5)3	5000
26. Snapinn & Knoke, 1989	Error rates & selection	2	10, 20	25	0, 1, 2, 3	100
27. Stam & Ragsdale, 1992	MP rules	2	3	25, 50, 100	$\sqrt{6.75}$= 2.598	50
28. Vlachonikolis, 1986	Error rates	2	1,3,5	50, 100, 200	.30, .50, .70*	50
29. Young, Marco & Odell, 1986	Dimension reduction	3	6	10,30,50 ,70,90	1.01, 1.31, 1.55	100

[1] Number of groups.
[2] Number of variables.
[3] Number of observations per group
[4] Mahalanobis distance; *: refers to expected error rate concept;
**: refers to a different concept of distance measure.
[5] Number of replications for every condition.
MP : Mathematical Programming approaches.

REFERENCES

Bayne, C. K., Beauchamp, J. J., Kane, V., & McCabe, G. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. Computational Statistics & Data Analysis, 1, 257-273.

Chatterjee, S., & Chatterjee, S. (1983). Estimation of misclassification probabilities by bootstrap methods. Communication in Statistics: Simulation and Computation, 12, 645-656.

Dorans, N. J. (1988). The shrunken generalized distance estimator of the actual error rate in discriminant analysis. Journal of Educational Statistics, 13, 63-74.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of The American Statistical Association, 78, 316-331.

Freed, N., & Glover, F. (1986). Evaluating alternative linear programming models to solve the two-group discriminant problem Decision Sciences, 17, 151-162.

Friedman, J. H. (1989). Regularized discriminant analysis. Journal of The American Statistical Association. 84, 165-175.

Fukunaga, K., & Kessell, D. (1971). Estimation of classification error. IEEE Transactions On Computers, 20, 1521-1527.

Ganesalingam, S., & McLachlan, G. J. (1980). Error rate estimation on the basis of posterior probabilities. Pattern Recognition, 12, 405-413.

Ganeshanandam, S., & Krzanowski, W. J. (1989). On selecting variables and assessing their performance in linear discriminant analysis. Australian Journal of Statistics, 31, 433-447.

Glick, N. (1978). Additive estimators for probabilities of correct classification. Pattern Recognition, 10, 211-222.

Greene, T., & Rayens, W. (1989). Partially pooled covariance matrix estimation in discriminant analysis. Communication in Statistics: Theory and Methods, 18, 3679-3702.

Hand, D.J. (1981). Discrimination and Classification, Wiley: Chichester.

Hora, S. C., & Wilcox, J. B. (1982). Estimation of error rates in several population discriminant analysis. Journal of Marketing Research, 19, 57-61.

Huberty, C. J., Wisenbaker, J. M., & Smith, J. C. (1987). Assessing predictive accuracy in discriminant analysis. Multivariate Behavioral Research, 22, 307-329.

Joachimsthaler, E. A., & Stam, A. (1988). Four approaches to the classification problem in discriminant analysis: an experimental study. Decision Sciences, 19, 322-333.

Joachimsthaler, E., & Stam, A. (1990). Mathematical programming approaches for the classification problem in two-group discriminant analysis. Multivariate Behavioral Research, 25, 427-454.

Koehler, G. L., & Erenguc, S. S., (1990). Minimizing Misclassifications in linear discriminant analysis. Decision Sciences, 21, 62-85.

Lachenbruch, P., & Mickey, M. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10, 1-11.

Lachenbruch, P. A., Sneeringer, C. & Revo, L. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. Communication in Statistics, 1,1-10.

Lahiff, M., & Whitcomb, K. (1990). Empirical influence function for misclassification rates in discriminant analysis. Communication in Statistics: Theory & Methods, 19, 2999-3009.

Lesaffre, E., & Willems, J. & Albert, (1989). Estimation of error rate in multiple group logistic discrimination. The approximate leaving-one-out method. Communication in Statistics: Theory & Methods, 18, 2989-3007.

McLachlan, G. J. (1980). The efficiency of Efron's "Bootstrap" approach applied to discriminant analysis. Journal of Statistical Computations & Simulations, 11, 273-279.

Moore, D. S., Whitsitt, S. J., & Landgrebe, D. A. (1976). Variance comparisons for unbiased estimators of probability of correct classification. IEEE Transactions On Information Theory, 22, 102-105.

Naylor, T. J., Balintfy, J. L., Burdick, D., & Chu, k. (1966). Computer simulation techniques. New York: Wiley.

Page, J. T. (1985). Error rate estimation in discriminant analysis. Technometrics, 27, 189-198.

Raveh, A., (1989). A nonmetric approach to linear discriminant analysis. Journal of The American Statistical Association. 84, 165-175.

Remme, J., Habbema, J. D., & Hermans (1980). A simulative comparison of linear, quadratic and kernel discrimination. Journal of Statistical Computations & Simulations, 11, 87-106.

Rubin, P. A. (1990). A comparison of linear programming and parametric approaches to the two-group discriminant problem. Decision Sciences, 21, 373-386.

Rubinstien, R. Y. (1981). Simulation and the Monte Carlo methods. New York: Wiley.

Sadek, R. F., & Huberty, C. J, (1992). On outlier influence in two-group classification analysis. Paper presented at the annual meetings of the American Educational Research Association (AERA) in San Francisco, April 20-24.

Snapinn, S. M., & Knoke, J. D. (1984). Classification error rate estimators evaluated by unconditional mean squared error. Technometrics, 26, 371-378.

Snapinn, S. M., & Knoke, J. D. (1985). An evaluation of smoothed classification error-rate estimators. Technometrics, 27, 199-206.

Snapinn, S. M., & Knoke, J. D. (1989). Estimation of error rates in discriminant analysis with selection of variables. Biometrics, 45, 289-299.

Stam, A., & Ragsdale, C. T. (1992). On the classification gap in MP based approaches to the discriminant problem. Naval Research Logistics.

Vlachonikolis, I. G. (1986). On the estimation of the expected probability of misclassification in discriminant analysis with mixed binary and continuous variables. Computers and Mathematics with Applications, 12, 187-198.

Young, M. D., Marco, R. V., & Odell, P. L. (1986). Dimension reduction for predictive discrimination. Computational Statistics & Data Analysis, 4, 243-25.